Rafael Nascimento

Predictive Model For Customer Behavior

Personal Presentation

Academic and Professional Background

Rafael Nascimento
Data Scientist, PMO

Rafael Nascimento has a postgraduate degree in Data Science and AI from FIA – Business School and a bachelor degree in Electrical Engineering from Polytechnic School of the University of São Paulo - Brazil, which is the most renowned education and research institution in South America and among the top hundred finest technology schools in the world according to the QS ranking.

His work experiences are in Data Science, Project Management and Supply Chain in the largest Pharmaceutical and Auto Parts Retailers in Brazil.

Since November 2020 he has been part of a Data Science initiative at B.Homy, which is a pioneering startup in property management for flexible rentals in Brazil.

Data Specialist

Planning

- Objective
- Context
- Dataset - ABT
- Main Variables

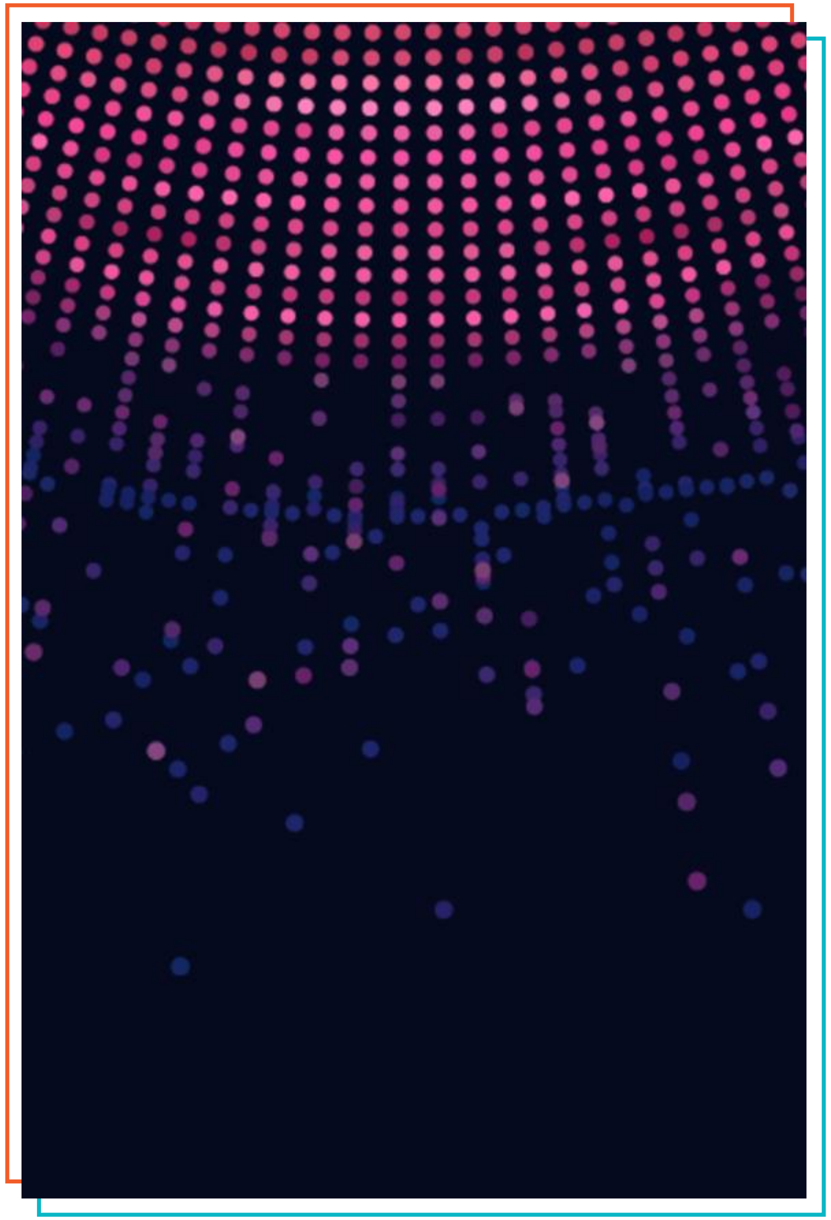
Planning

Objective

The present work aims to predict the customer behavior of a food retailer, indicating who in the customer database is most likely to purchase the offer of a new product.

In this sense, statistical and machine learning models will be developed over a dataset built from a pilot campaign.

Thus, the company will have a better understanding of how the variables affect the purchase and it will allow the company to maximize the profits of future campaigns.



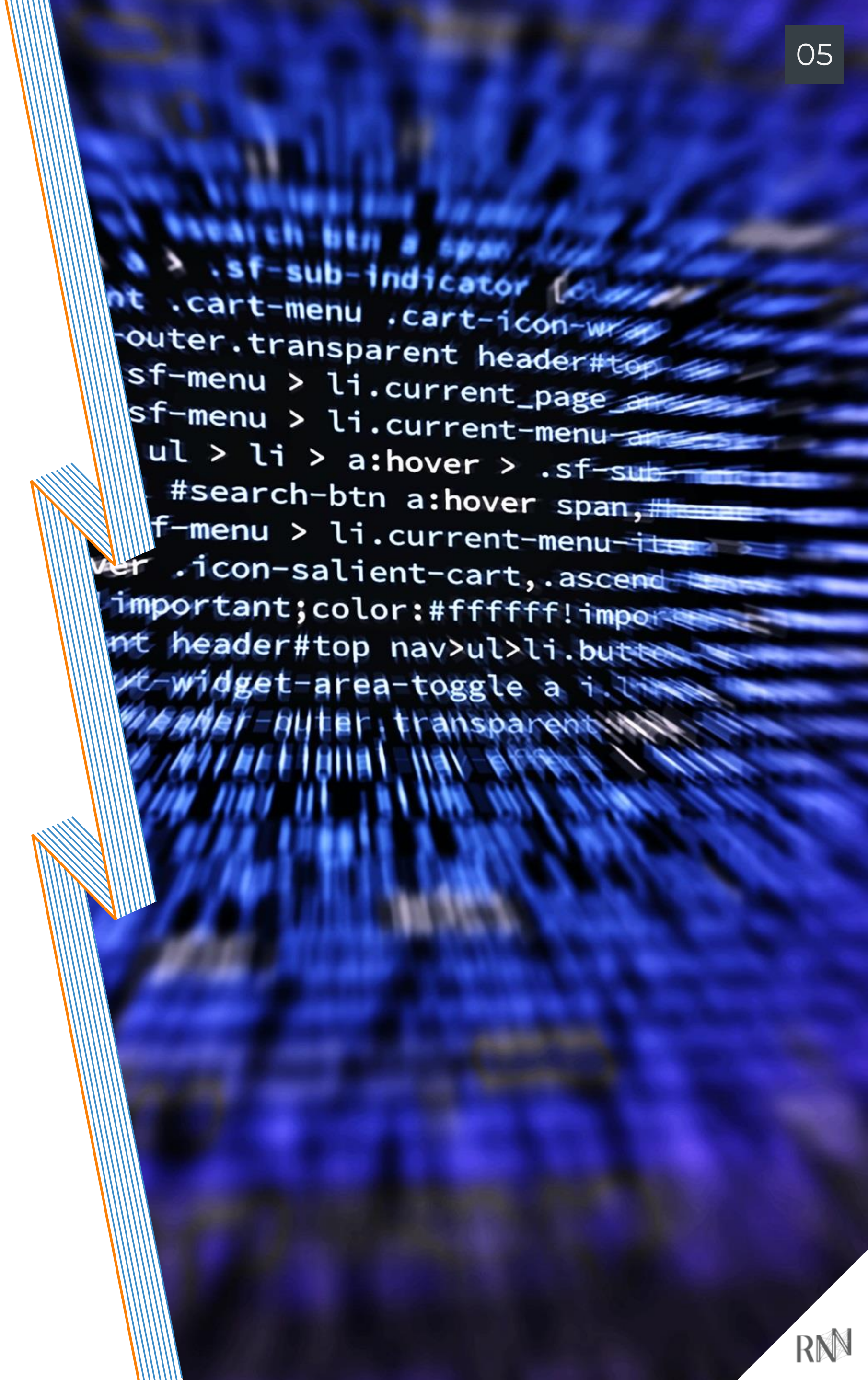
Planning

Context

A retailer with several thousand registered customers is implementing some initiatives to increase its revenues. As part of it, the marketing department is being pressured to improve the performance of the marketing campaigns.

A new campaign, the sixth, aims to sell a new product to the customer database. A pilot campaign involving 2.240 customers was carried out. The customers were randomly selected and contacted by phone, and the customers who bought the offer were labeled. The campaign had a negative profit, and the success rate was 15%.

The objective is to build a model that predicts customer behavior supporting the marketing department to apply it to the rest of the customers more wisely, thus leading to the maximization of profits.



Planning

Databases - ABT

Dataset - ABT

Shape:

2,240 rows x 27 columns

Variables:

26 features / 1 target (0 or 1)

The Analytical Base Table (ABT) contains demographic features of about 2,240 customers who were contacted by phone.

Additionally, it contains a flag for those customers who responded to the sixth campaign, by buying the product.

Planning

Main Variables

Registration Features

- Year of birth
- Customer’s level of education
- Customer’s marital status
- Number of children in the customer’s household
- Number of teenagers in the customer’s household
- Date of customer’s enrollment with the company

Customer Behavior Features

- Recency
- Complaints (0 or 1)
- Accepted campaign 1 (0 or 1)
- Accepted campaign 3 (0 or 1)
- Accepted campaign 4 (0 or 1)
- Accepted campaign 2 (0 or 1)
- Accepted campaign 5 (0 or 1)
- Purchases made in the stores
- Purchases made with discount
- Purchases made on the website

Financial Features

- Income

Target Variable

Response in the pilot campaign:
1 = customer accepted the offer
0 = customer didn’t accept the offer

Exploratory Data Analysis

- Summary Measures and Statistics
- Univariate / Bivariate Analysis
- Missing Values
- Outliers

Exploratory Data Analysis

Summary Measures and Statistics

● Customer Behavior Features

- Recency
- Complaints (0 or 1)
- Accepted campaign 1 (0 or 1)
- Accepted campaign 2 (0 or 1)
- Accepted campaign 3 (0 or 1)
- Accepted campaign 4 (0 or 1)
- Accepted campaign 5 (0 or 1)

Customer Behavior

- The number of visits to the company's website in the last month is left skewed (mean < median) where 1Q = 3, 2Q = 6 and 3Q = 7.
- The number of purchases made on the website is right skewed (mean > median) where skewness = 1.38.

	1st Campaign	2nd Campaign	3st Campaign	4st Campaign	5st Campaign
Sucess Rate	6.4%	1.4%	7.2%	7.4%	7.2%

Exploratory Data Analysis

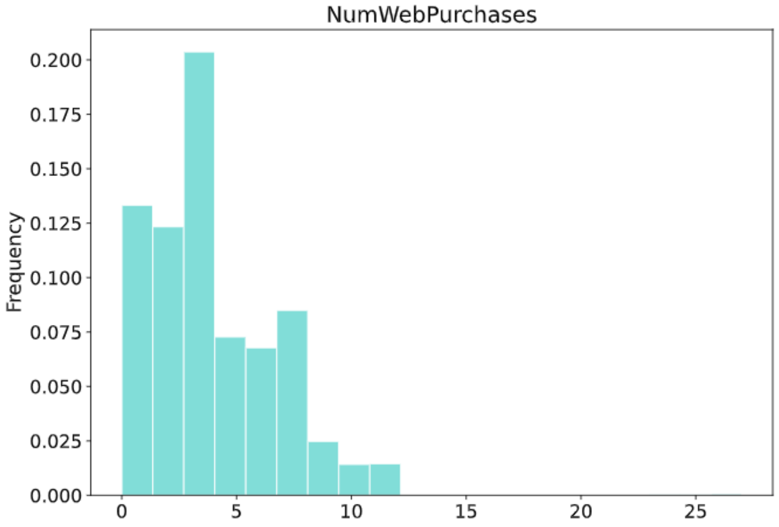
Summary Measures and Statistics

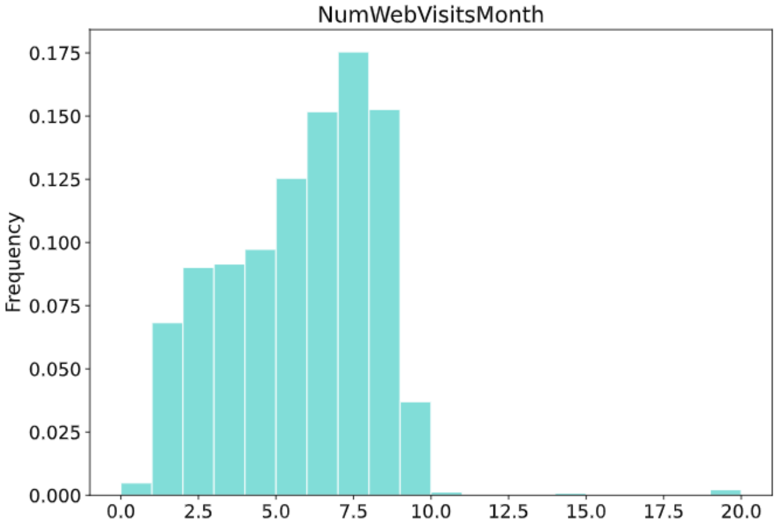
● **Customer Behavior Features**

- Nº of purchases made in the web site
- Nº of visits to the company's web site in the last month
- Nº of purchases made with discount

Customer Behavior

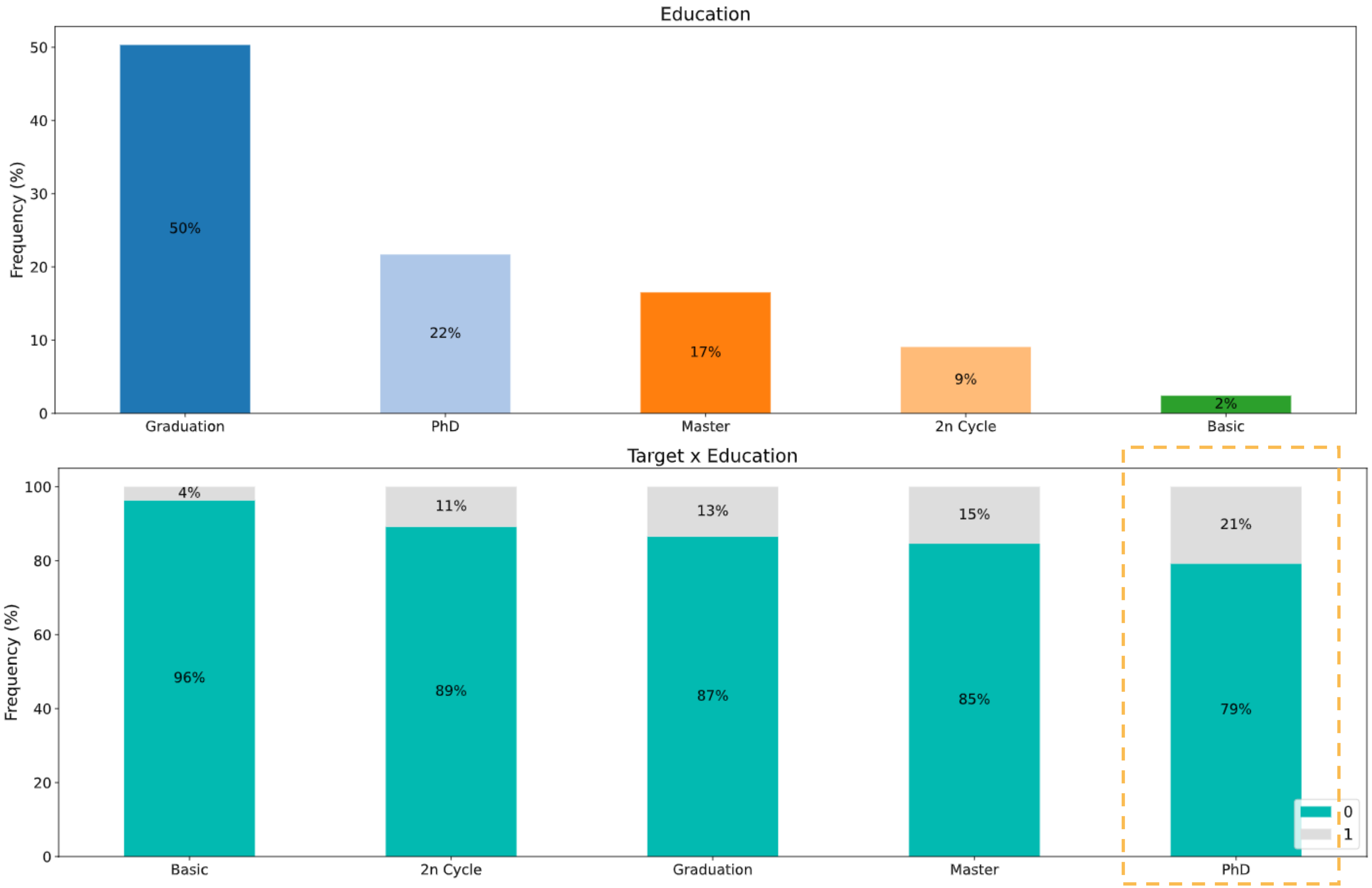
- The biggest recency registered in the database is 99 days and the distribution has a rectangular shape with a mean ≈ median of 49 days.
- Only 0.9% of the customers in the database have made any complaints in the last two years.





Exploratory Data Analysis

Univariate and Bivariate Analysis



Only 11% of the customers in the dataset don't have a college degree.

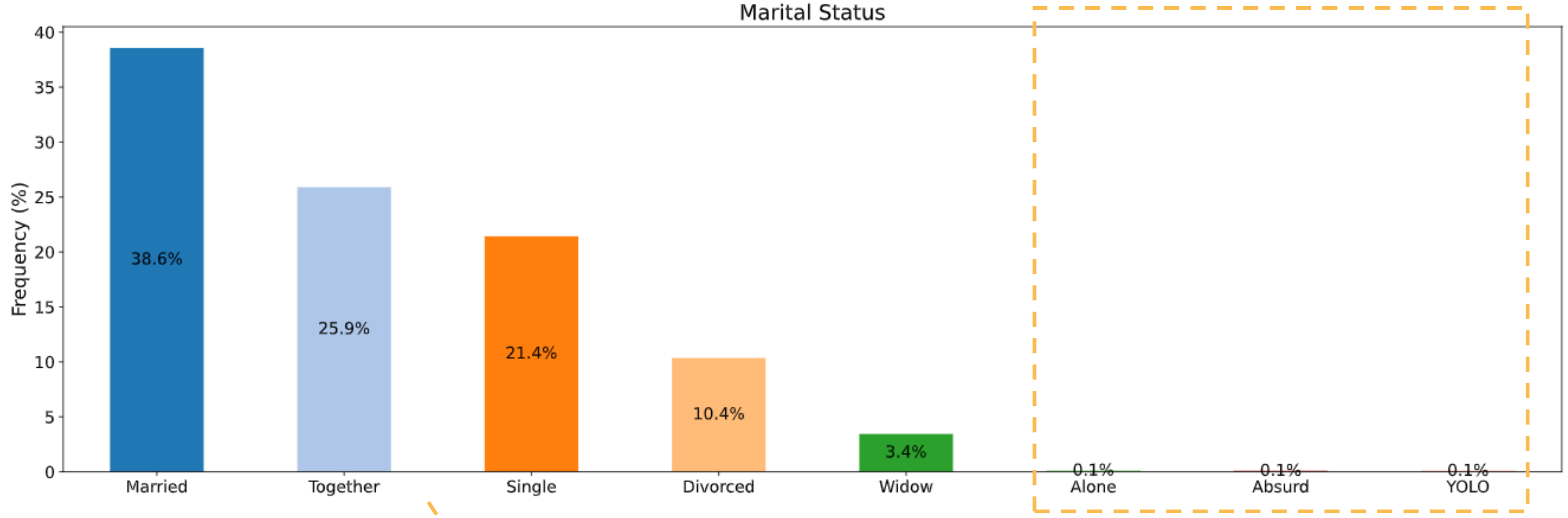
The pilot campaign had a success rate of 15%, but the success rate among the clients with Ph.D. was 21%.

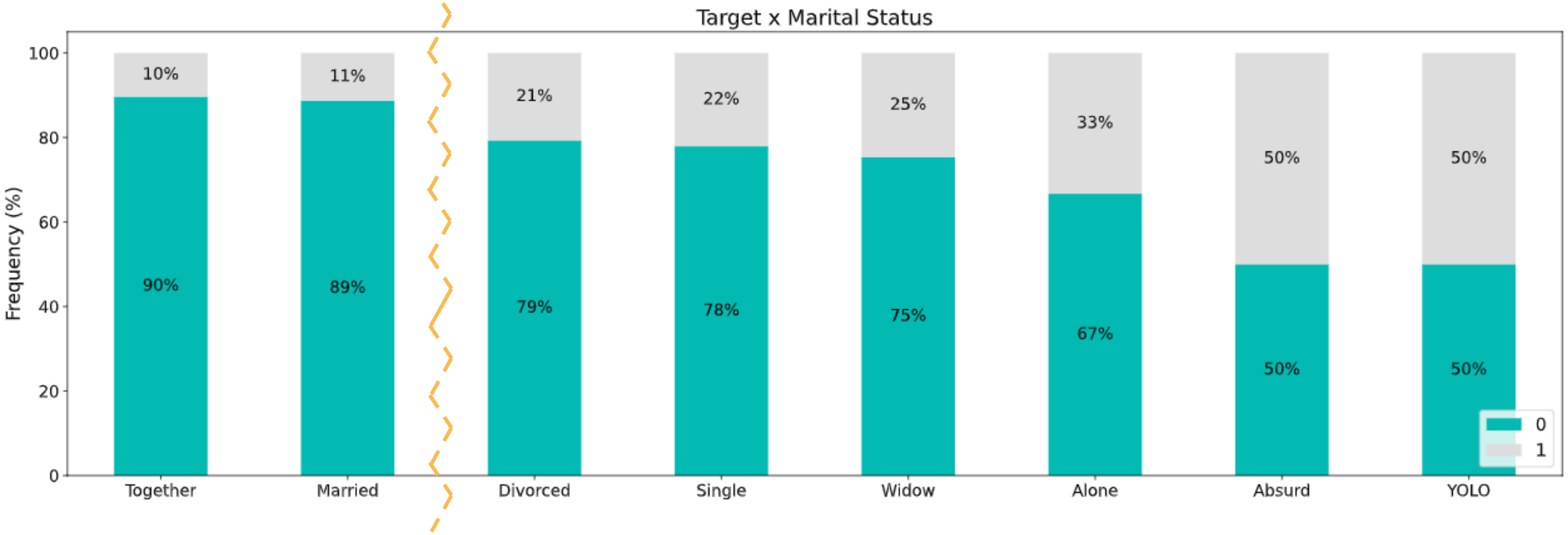
Target:

- 0 = the customer didn't accept the offer in the sixth campaign
- 1 = the customer accepted the offer in the sixth campaign

Exploratory Data Analysis

Univariate and Bivariate Analysis


Marital Status


Target x Marital Status

The categories **Alone**, **Absurd** and **YOLO** will be part of a new category called **Others.**

Clients who are not in a relationship are more likely to accept the offer.

Target:

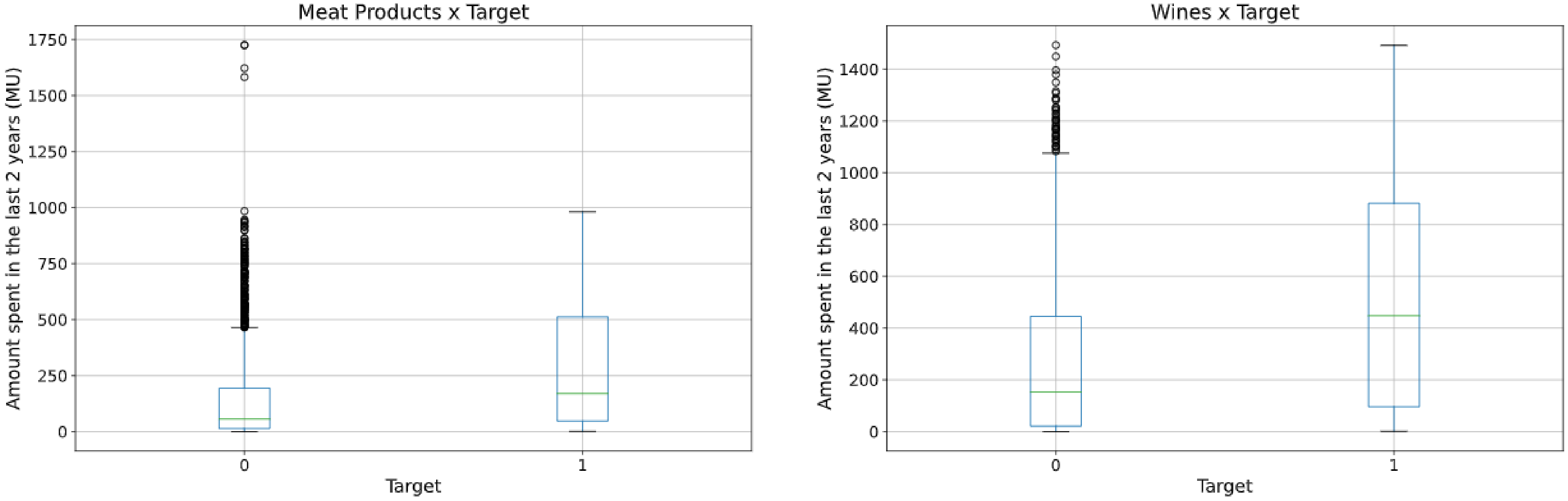
- 0 = the customer didn't accept the offer in the sixth campaign
- 1 = the customer accepted the offer in the sixth campaign

Exploratory Data Analysis

Bivariate Analysis

Costumer Behavior

Clients who accepted the offer spent more money on meat/wine products in the last two years when compared to those that didn‘t.



Target:

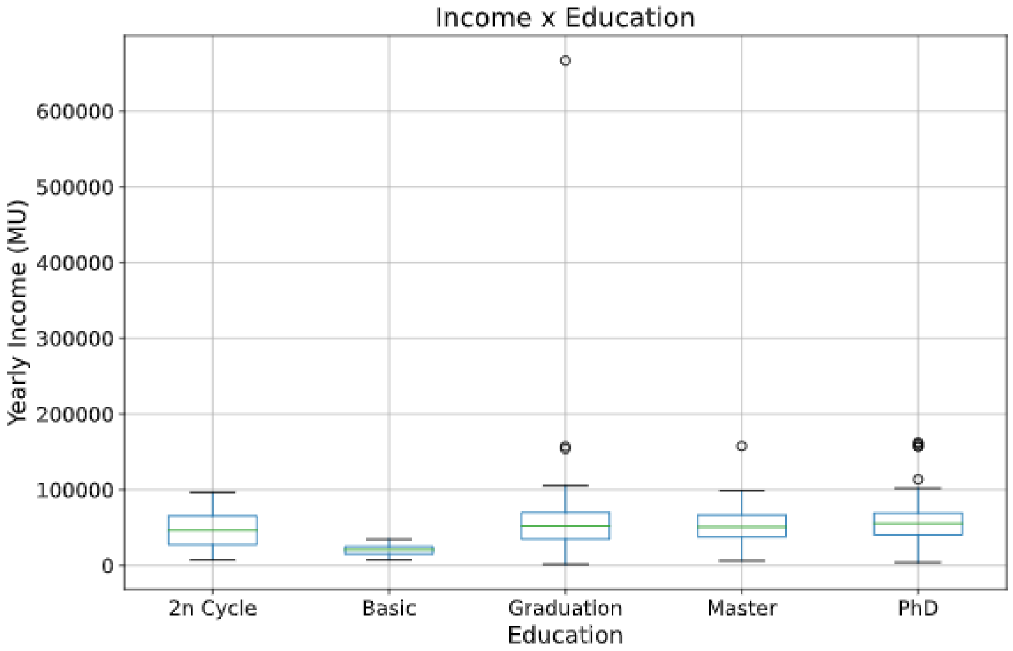
- 0 = the customer didn't accept the offer in the sixth campaign
- 1 = the customer accepted the offer in the sixth campaign

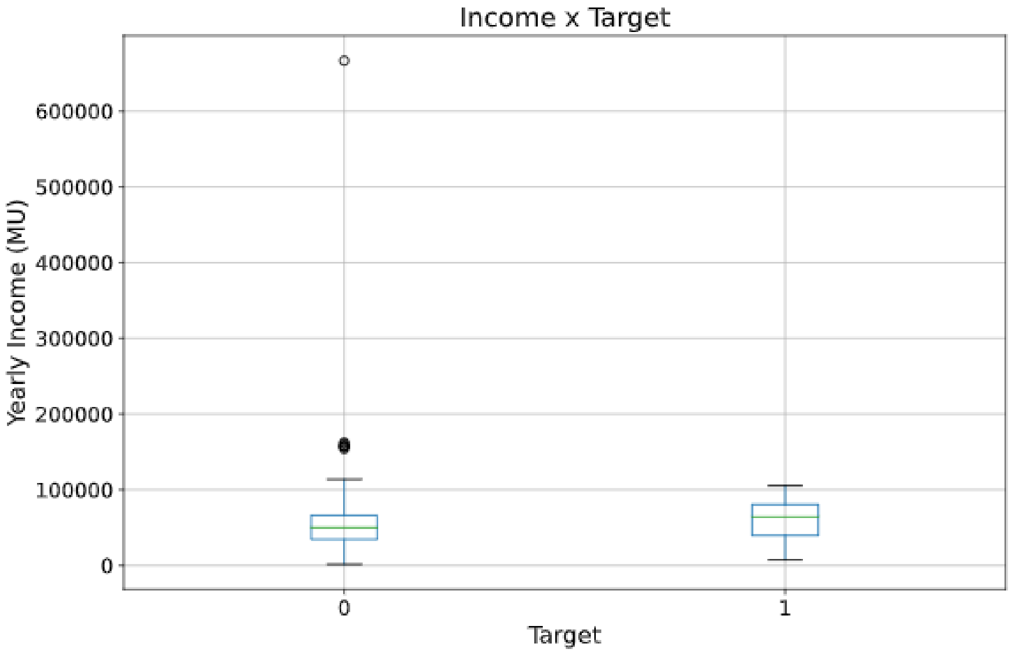
Exploratory Data Analysis

Bivariate Analysis

Education, Income and Target

The first boxplot brings that clients who have at least a college degree have a more equal income level. The second boxplot shows that the average income of the group that accepted the offer is higher than the group that didn't.





Target:

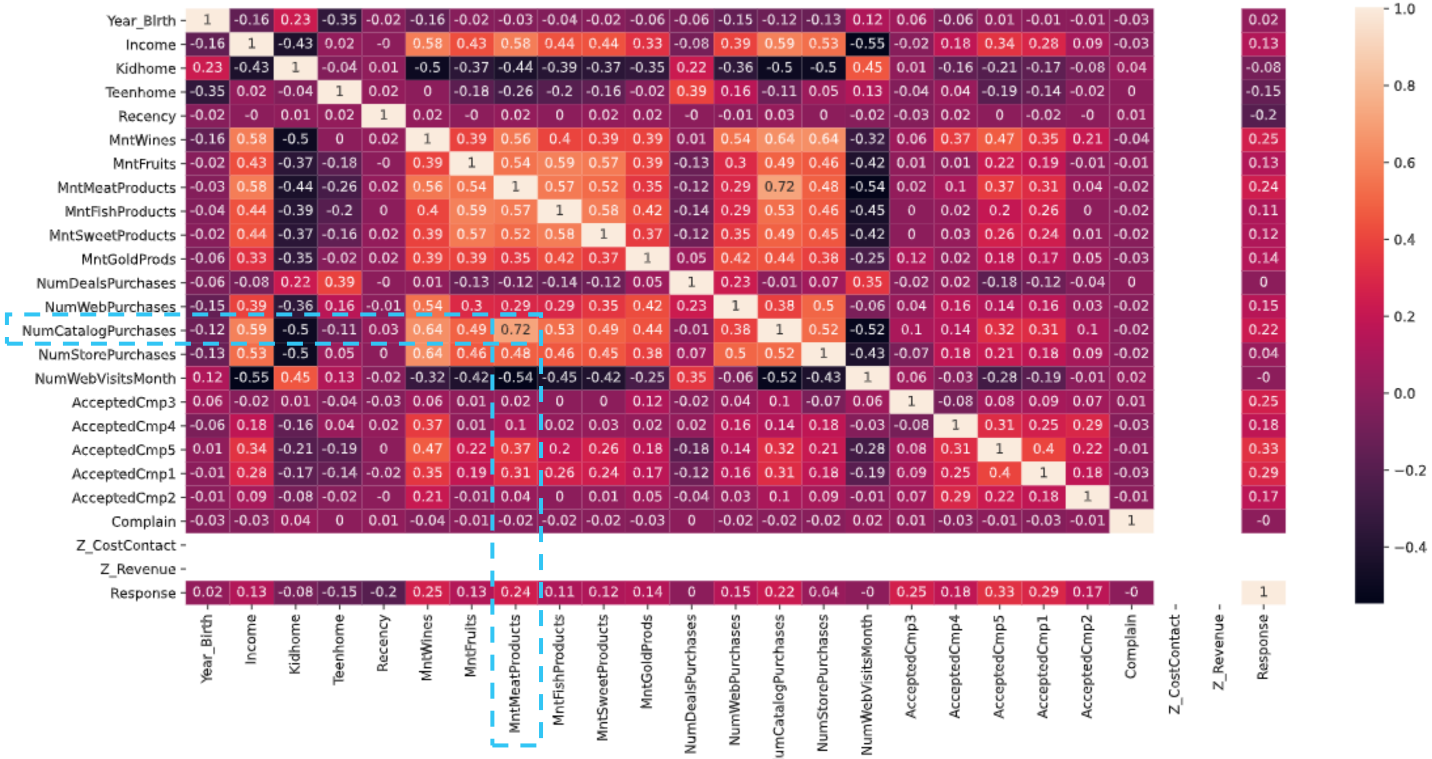
- 0 = the customer didn't accept the offer in the sixth campaign
- 1 = the customer accepted the offer in the sixth campaign

Exploratory Data Analysis

Bivariate Analysis

Correlogram

The amount spent on meat products and the number of purchases made using catalogs have the highest correlation among all variables in the database.



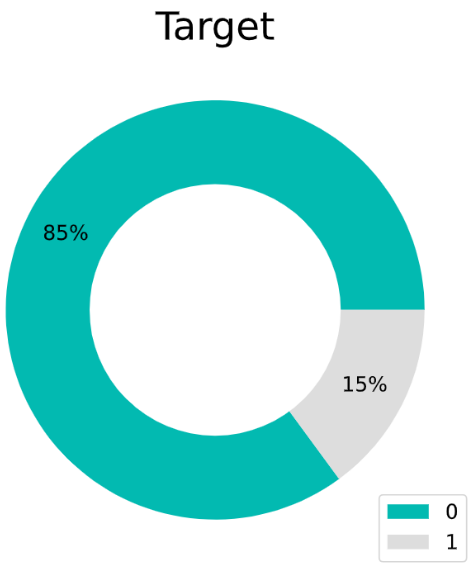
Considering the threshold of 70%, it is worth paying close attention to multicollinearity issues when modeling the Logistic Regression.

Exploratory Data Analysis

Univariate Analysis – Target Variable

Pilot Campaign

Of the 2,240 customers in the database, 15% accepted the offer in the pilot campaign.



Unbalanced Classes

- The database is unbalanced, but since there are only 2,240 clients, a rebalancing by under-sampling the majority class may not result in a better performance.
- It will be tested in further modeling considering the limitations mentioned.

Modeling with Traditional Statistics

- Modeling Methodology
- Decision Tree x Logistic Regression
- Logistic Regression - Threshold Issue
- Logistic Regression - Feature Importance
- Summary Results

Modeling with Traditional Statistics

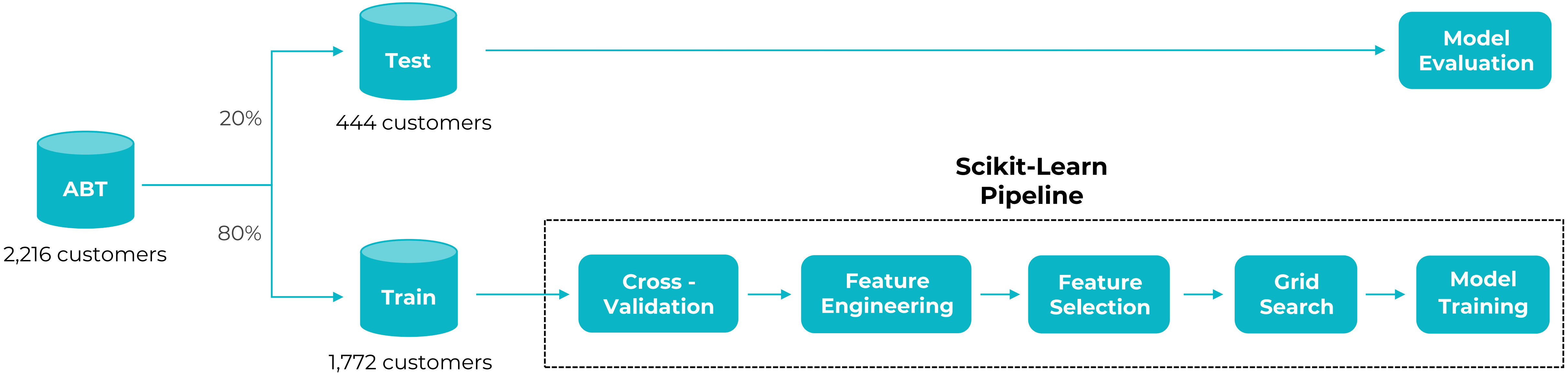
Modeling Methodology

Training and Test Datasets

For the training and testing datasets, the 80/20 ratio was used:

- 80% random for training and validation → 1,772 customers
- 20% random for testing → 444 customers

Modeling Workflow



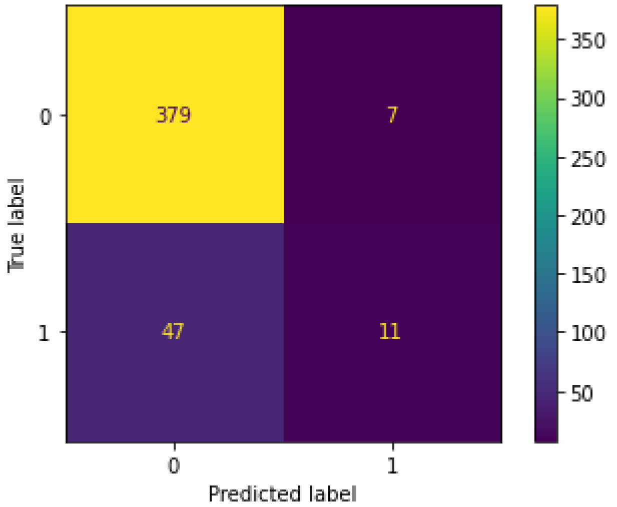
Modeling with Traditional Statistics

Decision Tree x Logistic Regression

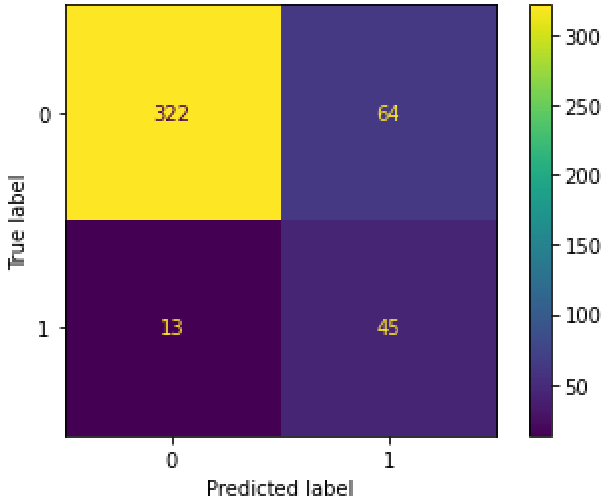
Assumptions

- The threshold used in the both confusion matrix was 0.5
- Unit cost = 3 MU
- Unit sales revenue = 11 MU

Decision Tree – Confusion Matrix



Logistic Regression – Confusion Matrix



Decision Tree - Results

		Train	Test
Precision	=	73.8%	61.1%
Recall	=	33.5%	19.0%
F1	=	45.6%	28.9%
AUC	=	0.80	**0.77**
Profit	=	11*11 – 3*(11+7) = **78 MU**	

Logistic Regression - Results

		Train	Test
Precision	=	40.1%	41.3%
Recall	=	75.3%	77.6%
F1	=	52.3%	53.9%
AUC	=	0.87	**0.90**
Profit	=	11*45 – 3*(45+64) = **168 MU**	

Modeling with Traditional Statistics

Logistic Regression - Threshold Issue

Logistic Regression – AUC = 0.9

Threshold	TP	FP	FN	Precision	Recall	Profit (MU)	Return	% Costumers Reached
0.1	58	290	0	17%	100%	-406	-39%	78%
0.2	57	206	1	22%	98%	-162	-21%	59%
0.3	54	140	4	28%	93%	12	2%	44%
0.4	50	92	8	35%	86%	124	29%	32%
0.5	45	64	13	41%	78%	168	51%	25%
0.6	42	48	16	47%	72%	192	71%	20%
0.7	35	25	23	**58%**	**60%**	**205**	**114%**	**14%**
0.8	30	15	28	67%	52%	195	144%	10%
0.9	18	7	40	**72%**	**31%**	**123**	**164%**	**6%**

✕

Random – AUC = 0.5

Initial Scenerio
Customers: 2,240
Profit: **-3,024 MU**
Return: **-45%**
Precision: **15%**
% Customers Reached: **100%**

Which is the most suitable threshold?

There is no absolute right answer to this question. It will vary depending on the business situation.

A company with a large base of customers and a low budget probably would recommend its marketing team to contact only customers scored with a probability higher than 0.9.

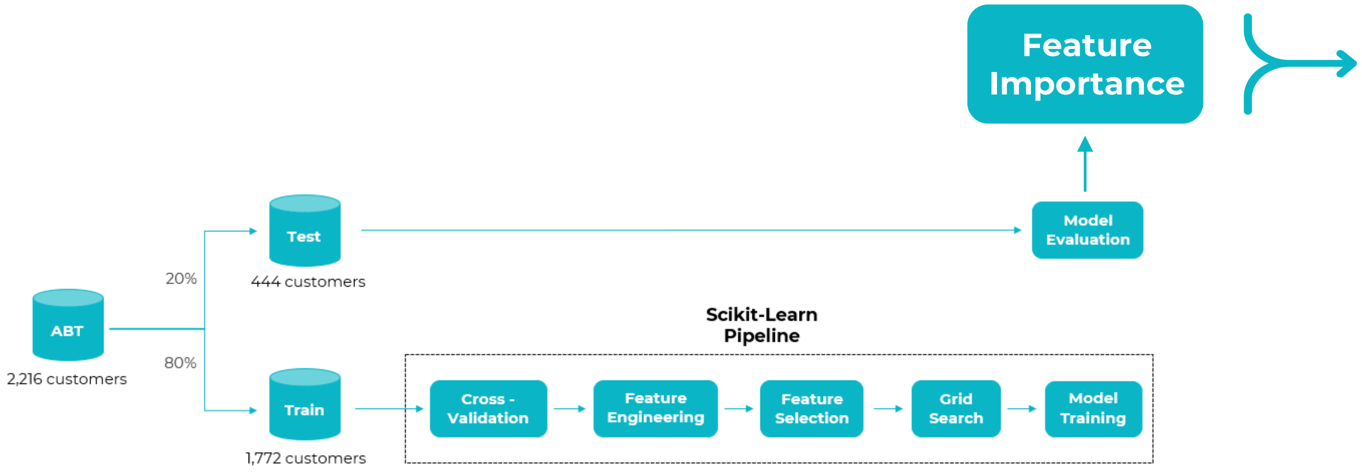
A company with the risk of getting its level of spoiled stocks out of control possibly would be willing to finance a marketing campaign to avoid wasting an expensive stock. In this sense, perhaps the threshold of 0.2 or 0.3 would make more sense.

There are endless business possibilities to further illustrate. With the table above, a business manager would be able to make a decision based on data and supported by a data scientist considering their business needs.

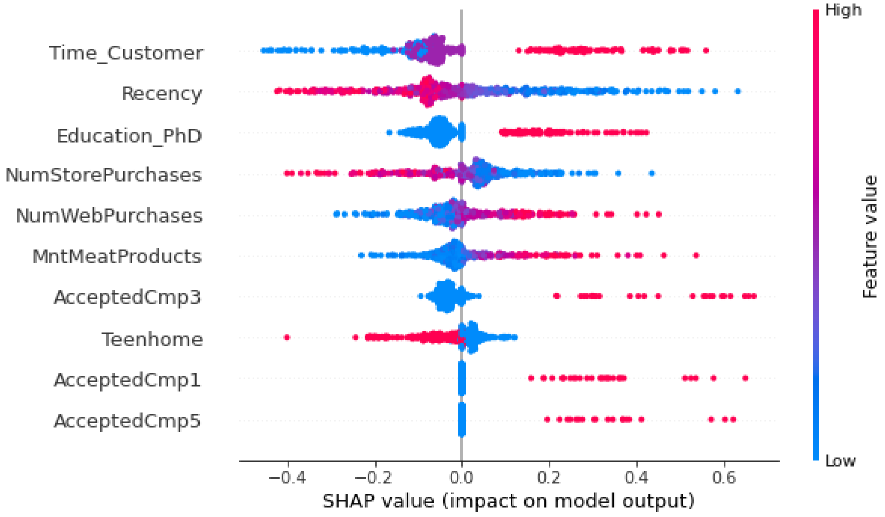
Modeling with Traditional Statistics

Logistic Regression - Feature Importance

Logistic Regression – Feature Importance



Shap Values – Explaining Machine Learning Models



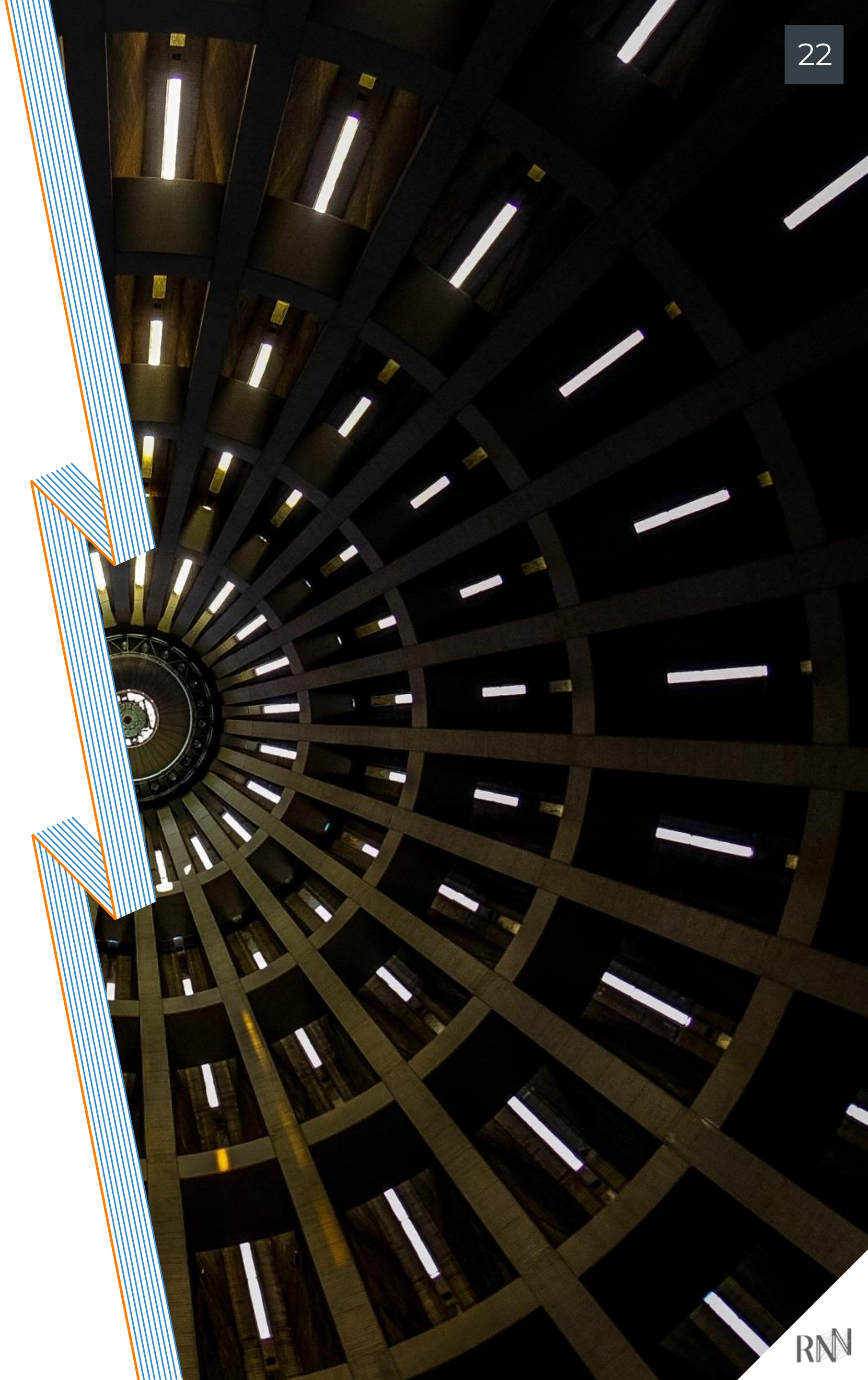
Feature Importance Analysis

- Loyal customers with lower recency impact the model output positively.
- As verified in the exploratory data analysis, customers with Ph.D. are more likely to accept the offer in the campaign.
- Customers more likely to accept the offer spent more money on meat products in the last two years and have the behavior to buy online instead of in the stores.
- Customers with kids impact negatively the model output.
- Customers who accepted the first, third, and fifth campaigns impact positively the model output.

Modeling with Traditional Statistics

Summary Results

- The dataset used for modeling with traditional statistics has a small number of observations and a relatively large number of predictors. Moreover, the dataset is imbalanced.
- Models trained on a small number of observations and a large number of predictors tend to overfit and produce inaccurate results. The number of predictors dropped from 26 to 10 after feature selection.
- Logistic Regression performed better than the Decision Tree.
- The Scenario 3 (threshold = 0.7) - or if the marketing team only contact customers scored with probability higher than 0.7 - return the maximum absolute profit of 205 MU.
- A rebalancing in the dataset didn't bring a better performance in the Logistic Regression (small dataset issue).
- Top 3 strongest predictors: customer retention (time), recency and customer with Ph.D.



Modeling with AI

- Modeling Methodology
- PyCaret - Machine Learning Algorithms
- LGBM x CatBoost
- CatBoost - Feature Importance
- Summary Results

Modeling with AI

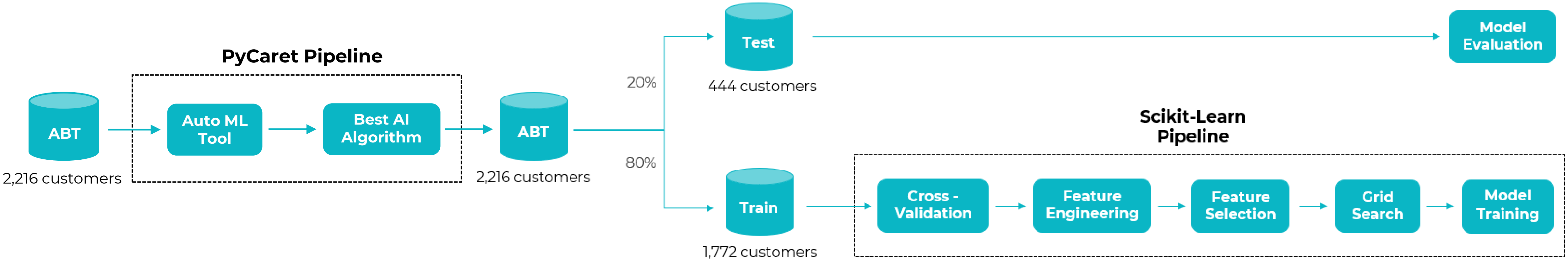
Modeling Methodology

Training and Test Datasets

For the training and testing datasets, the 80/20 ratio was used:

- 80% random for training and validation → 1,772 customers
- 20% random for testing → 444 customers

Modeling Workflow



Modeling with AI

PyCaret – Machine Learning Algorithms

PyCaret

PyCaret is an auto machine learning tool, in which it is possible to explore a set of models all at once with low effort.

PyCaret - Results

Models	AUC
CatBoost Classifier	0.89
Light Gradient Boosting Machine	0.89
Gradient Boosting Classifier	0.88
Logistic Regression	0.88
Extreme Gradient Boosting	0.88
Linear Discriminant Analysis	0.88
Random Forest Classifier	0.86
Ada Boost Classifier	0.86
Extra Trees Classifier	0.85
Naive Bayes	0.77
K Neighbors Classifier	0.75
Decision Tree Classifier	0.68
Quadratic Discriminant Analysis	0.51

- The CatBoost Classifier and the LGBM presented the highest AUC among 13 models.
- Next steps: go further in the modeling with these two algorithms using the framework Scikit-Learn.

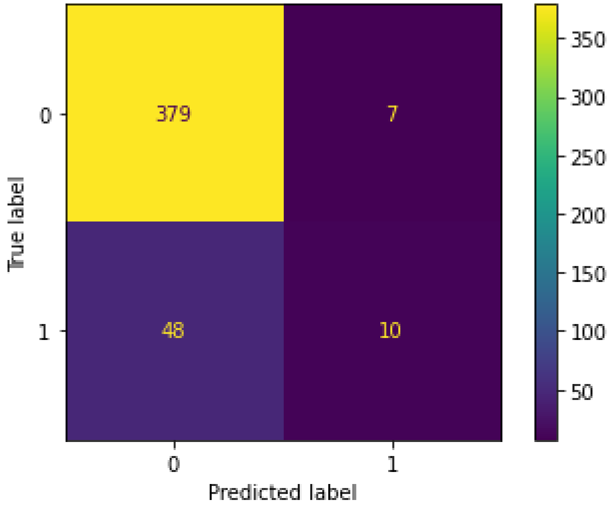
Modeling with AI

LGBM x CatBoost

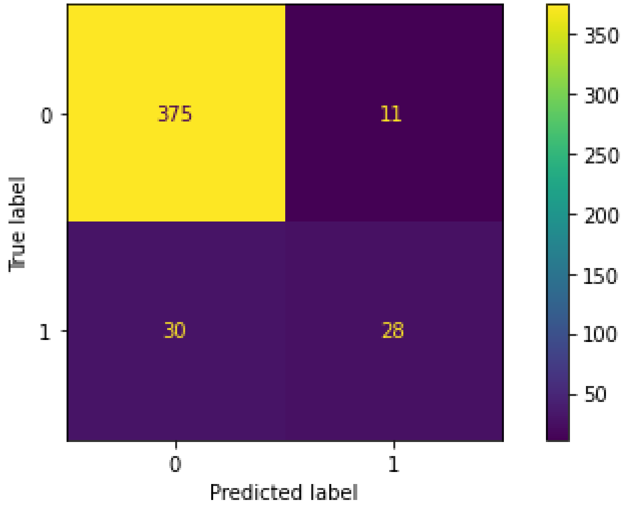
Assumptions

- The threshold used in the both confusion matrix was 0.5
- Unit cost = 3 MU
- Unit sales revenue = 11 MU

LGBM – Confusion Matrix



CatBoost – Confusion Matrix



LGBM - Results

		Train	Test
Precision	=	72.6%	58.8%
Recall	=	30.5%	17.2%
F1	=	42.7%	26.7%
AUC	=	0.85	**0.86**
Profit	=	11*10 – 3*(10+7) = **59 MU**	

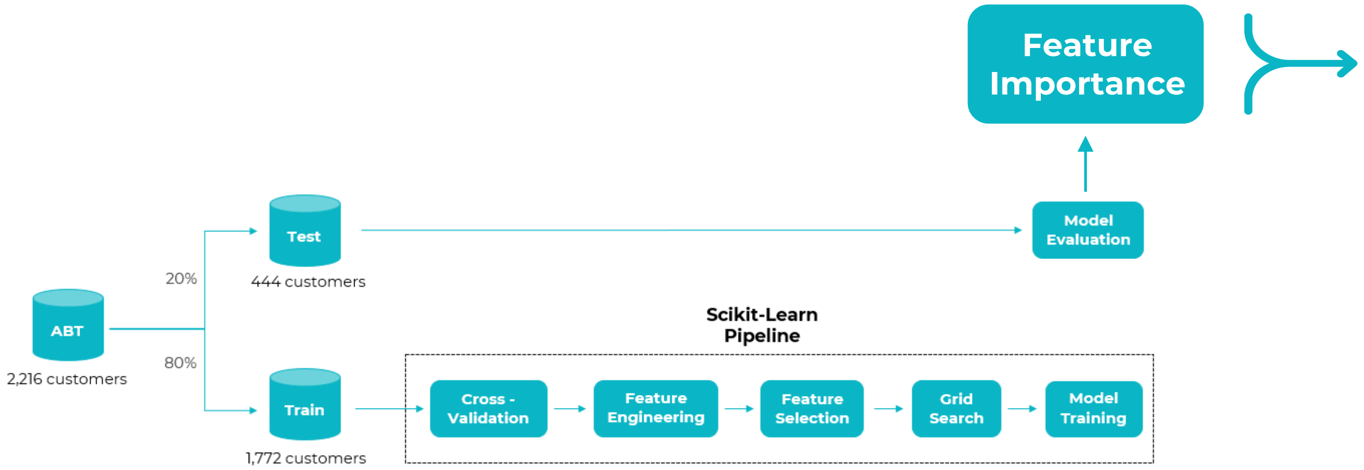
CatBoost - Results

		Train	Test
Precision	=	66.9%	71.8%
Recall	=	34.9%	48.3%
F1	=	45.5%	57.7%
AUC	=	0.87	**0.90**
Profit	=	11*28 – 3*(28+11) = **191 MU**	

Modeling with AI

CatBoost - Feature Importance

CatBoost – Feature Importance



Feature Importance

Model Evaluation

Test

444 customers

20%

ABT

2,216 customers

80%

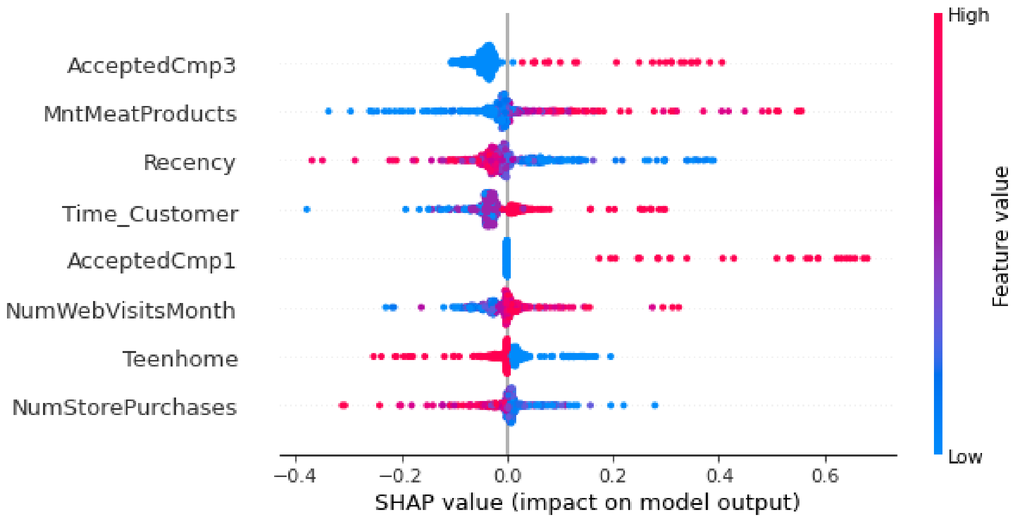
Train

1,772 customers

Scikit-Learn Pipeline

Cross - Validation → Feature Engineering → Feature Selection → Grid Search → Model Training

Shap Values – Explaining Machine Learning Models



Feature Importance Analysis

- Loyal customers with lower recency impact the model output positively.
- Customers more likely to accept the offer spent more money on meat products in the last two years and have the behavior to buy online instead of in the stores.
- Customers with no teenagers in the household impact positively the model output.
- Customers who accepted the first and third campaigns impact positively the model output.

Modeling with AI

Summary Results

- The dataset used for modeling with traditional statistics has a small number of observations and a relatively large number of predictors. Moreover, the dataset is imbalanced.
- Models trained on a small number of observations and a large number of predictors tend to overfit and produce inaccurate results. The number of predictors dropped from 26 to 8 after feature selection.
- CatBoost Classifier performed better than the LGBM.
- The threshold = 0.4 - or if the marketing team only contact customers scored with probability higher than 0.4 - return the maximum absolute profit of 206 MU.
- Even with a complex model such CatBoost, the generalization in the test dataset was great, evidencing that the overfitting was avoided.
- A rebalancing in the dataset didn't bring a better performance in the CatBoost Classifier (small dataset issue).
- Top 3 strongest predictors: customer accepted campaign 3, recency and the money spent on meat products in the last two years.

AI x Traditional Statistics

- CatBoost x Logistic Regression - Threshold Issue
- CatBoost x Logistic Regression - Performances and Characteristics
- CatBoost x Logistic Regression - Feature Importance

AI x Traditional Statistics

CatBoost x Logistic Regression - Threshold Issue

CatBoost – AUC = 0.9

Threshold	TP	FP	FN	Profit (MU)	Return	% Costumers Reached
0.1	52	119	6	59	12%	39%
0.2	45	60	13	180	57%	24%
0.3	38	29	20	217	108%	15%
0.4	31	14	27	**206**	**153%**	**10%**
0.5	28	11	30	**191**	**163%**	**9%**
0.6	15	6	43	102	162%	5%
0.7	3	2	55	18	120%	1%
0.8	2	1	56	13	144%	0.7%
0.9	1	1	57	5	83%	0.5%

Logistic Regression – AUC = 0.9

Threshold	TP	FP	FN	Profit (MU)	Return	% Costumers Reached
0.1	58	290	0	-406	-39%	78%
0.2	57	206	1	-162	-21%	59%
0.3	54	140	4	12	2%	44%
0.4	50	92	8	124	29%	32%
0.5	45	64	13	168	51%	25%
0.6	42	48	16	192	71%	20%
0.7	35	25	23	**205**	**114%**	**14%**
0.8	30	15	28	195	144%	10%
0.9	18	7	40	**123**	**164%**	**6%**

CatBoost x Logistic Regression Analysis

- Comparing the tables above is possible to verify that both models are able to deliver the same level of profits and returns.
- The main difference between the models is that CatBoost is more criterion than the Logistic Regression when scoring customers with higher probabilities.
- In other words, the CatBoost is more strict in scoring the customers, which characteristic possibly would be suitable for companies with a large number of clients or a low budget.

AI x Traditional Statistics

CatBoost x Logistic Regression - Performances and Characteristics

CatBoost

206 MU

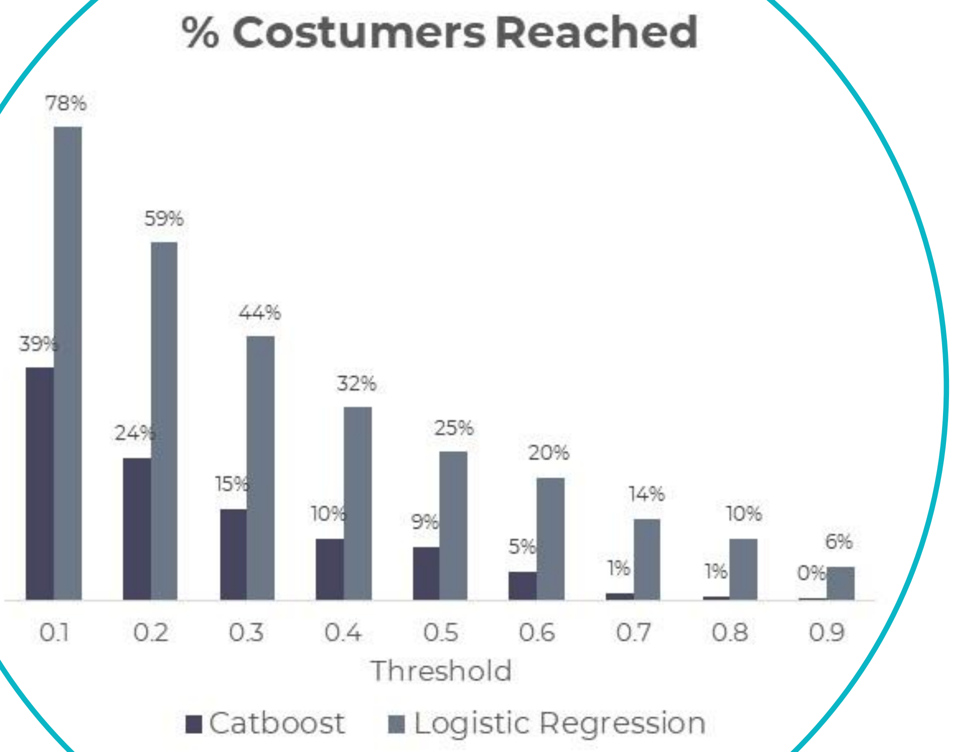
Profit

Threshold = 0.4

163%

Return

Threshold = 0.5



Logistic Regression

205 MU

Profit

Threshold = 0.7

164%

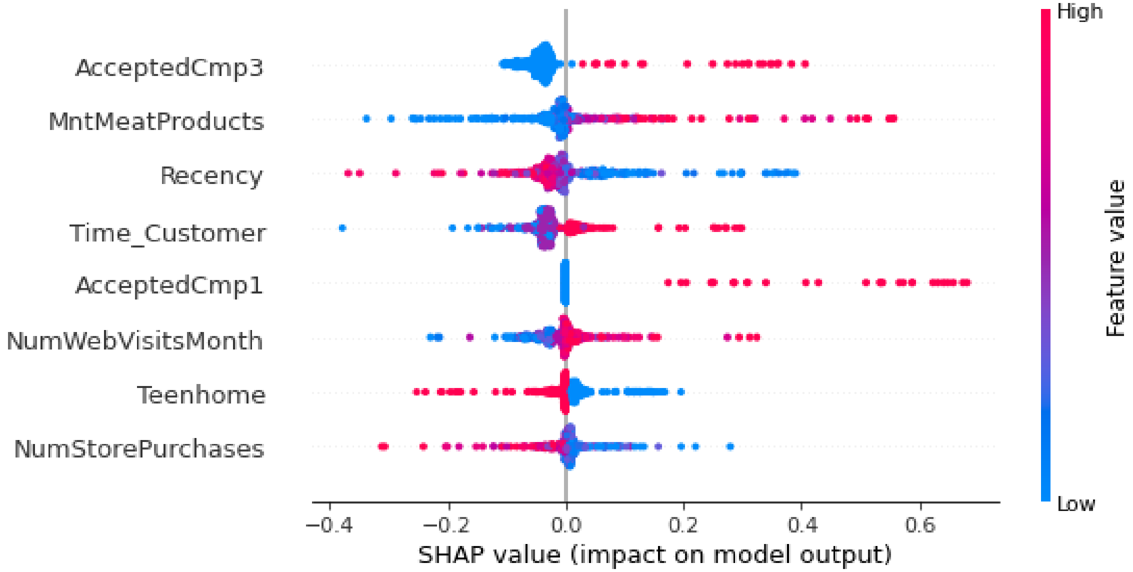
Return

Threshold = 0.9

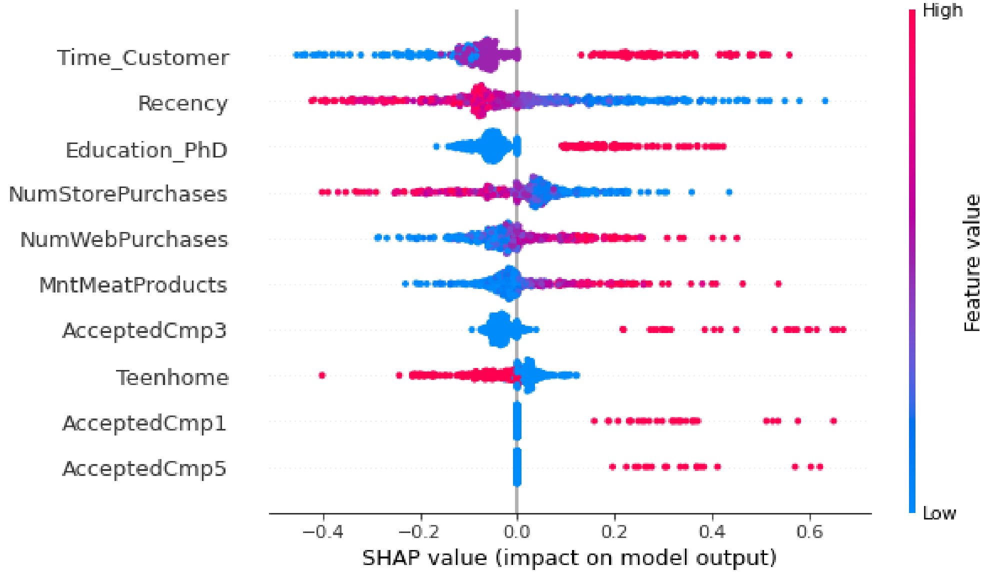
AI x Traditional Statistics

CatBoost x Logistic Regression - Feature Importance

CatBoost – Shap Values



Logistic Regression – Shap Values



Feature Importance Analysis

- The most interesting difference between the models regarding feature importance is the absence of the variable Education_PhD in the CatBoost Shap Values outputs.

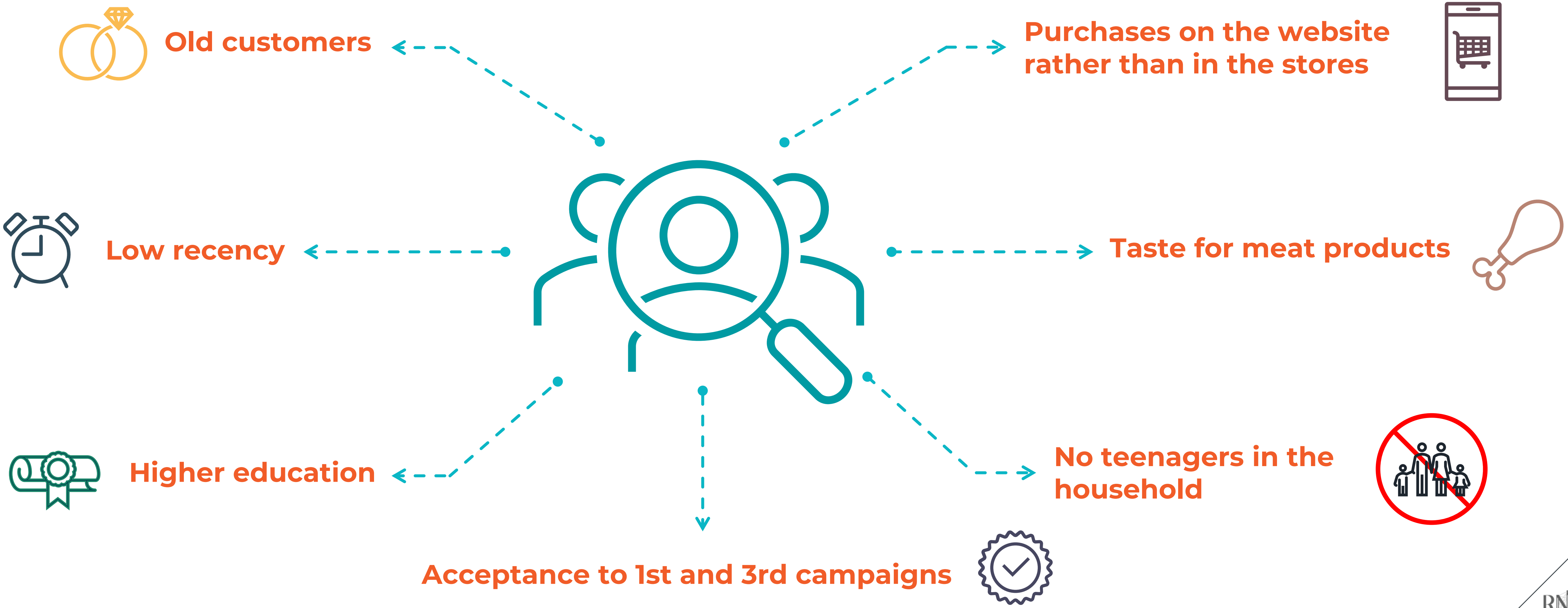
Business Gains and Insights

- The Customer’s Profile the CMO Search For
- Scenario Comparison - Results

Business Gains and Insights

The Customer’s Profile the CMO Search For

Characteristic Features of Customers Most Likely to Purchase the Offer



Old customers

Purchases on the website rather than in the stores

Low recency

Taste for meat products

Higher education

No teenagers in the household

Acceptance to 1st and 3rd campaigns

Business Gains And Insights

Scenario Comparison - Results

Initial Situation

- The marketing department is being pressured to improve the performance of the marketing campaigns.
- A new campaign, the sixth, aims to sell a new product to the customer database.
- A pilot campaign involving 2,240 customers was carried out.
- The customers were randomly selected and contacted by phone.

Pilot Campaign - Results

x Sucess Rate: **15%**

x Profit: **-3,024 MU**

x Return: **-45%**

x Customers Reached: **100% of 2,240**

Final Situation

- A statistical and machine learning model were developed over a dataset built from a pilot campaign.
- The CMO now has a quantitative approach to maximize the profit of the next marketing campaign.
- The marketing department got a gain of awareness about the customer profile who are willing to buy the new product.

ML Model – Results (Test Dataset)

✓ Sucess Rate: **72%**

✓ Profit: **+191 MU**

✓ Return: **+163%**

✓ Customers Reached: **9% of 444**

RNN

End

Thank You!